INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Stability Analysis of GI/G/c/K Retrial Queue with Constant Retrial Rate

Konstantin Avrachenkov — Evsey Morozov

N° 7335

July 2010

Thème COM

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7335--FR+ENG

Stability Analysis of GI/G/c/K Retrial Queue with Constant Retrial Rate

Konstantin Avrachenkov* , Evsey Morozov†

Thème COM — Systèmes communicants
Projet Maestro

Rapport de recherche n° 7335 — July 2010 — 16 pages

Abstract: We consider a GI/G/c/K-type retrial queueing system with constant retrial rate. The system consists of a primary queue and an orbit queue. The primary queue has c identical servers and can accommodate the maximal number of K jobs. If a newly arriving job finds the full primary queue, it joins the orbit. The original primary jobs arrive to the system according to a renewal process. The jobs have general i.i.d. service times. A job in front of the orbit queue retries to enter the primary queue after an exponentially distributed time independent of the orbit queue length. Telephone exchange systems, Medium Access Protocols and short TCP transfers are just some applications of the proposed queueing system. For this system we establish minimal sufficient stability conditions. Our model is very general. In addition, to the known particular cases (e.g., M/G/1/1 or M/M/c/c systems), the proposed model covers as particular cases the deterministic service model and the Erlang model with constant retrial rate. The latter particular cases have not been considered in the past. The obtained stability conditions have clear probabilistic interpretation.

Key-words: Retrial Queue, Constant Retrial Rate, Renewal Process, Regeneration, Stochastic Stability

* INRIA Sophia Antipolis-Méditerranée, France, k.avrachenkov@sophia.inria.fr
† Institute of Applied Mathematical Research, Karelian Research Centre RAS and Petrozavodsk University, Russia, emorozov@karelia.ru

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

L'analyse de la stabilité de la file d'attente de type GI/G/c/K avec des clients qui reviennent à un taux constant

Résumé : On considère une file d'attente de type GI/G/c/K avec des clients qui reviennent à un taux constant. Le système se compose d'une file d'attente primaire et une file d'attente orbite. La file d'attente primaire a c serveurs identiques et peut accueillir le nombre maximal de K clients. Si un arrivé trouve la file d'attente primaire pleine, il rejoint l'orbite. Les clients qui entrent dans le système pour la première fois arrivent selon un processus de renouvellement. Les clients ont un temps de service générale iid. Les clients dans la file d'attente orbite essaient d'entrer dans la file d'attente primaire après un temps avec une distribution exponentielle indépendante de la longueur de la file d'attente orbite. Les commutateurs téléphoniques, le contrôle d'accès au support, et les courte transferts TCP sont quelques-unes des applications de le système étudié. Pour ce système, nous établissons les conditions de stabilité suffisantes. Notre modèle est très général. En plus des cas particuliers (par exemple, M/G/1/1 ou M/M/c/c), le modèle proposé couvre les cas particuliers du modèle de service déterministe et le modèle Erlang avec des clients qui reviennent. Les derniers cas particuliers n'ont pas été considéré dans le passé. Les conditions de stabilité obtenus ont une interprétation probabiliste tres claire.

Mots-clés : File d'Attente avec des Clients qui Reviennent, Stabilité Stochastique, Processus de Renouvellement, Regeneration

1 Introduction

We consider a GI/G/c/K-type retrial queueing system. The system consists of a primary queue and an orbit queue. The primary queue has c identical servers and can accommodate the maximal number of K jobs. If a newly arriving job finds the full system, it joins the orbit. The original primary jobs arrive to the system according to a renewal process with rate λ. We denote the arrival times of the original primary jobs by t_n and we denote the interarrival times by $\tau_n = t_{n+1} - t_n$, $n \geq 1$, with generic element τ. Without loss of generality we assume that $t_1 = 0$. The jobs have general i.i.d. service times $\{S_n^{(1)}, n \geq 1\}$ with service rate μ and generic element $S^{(1)}$. Retrial times $\{S_n^{(2)}, n \geq 1\}$ are i.i.d. exponential with (generic) service time $S^{(2)}$ with rate μ_0 and independent of the orbit size, provided it is positive. Such a model is referred to as a retrial model with *constant retrial rate*. It then follows that the orbit can be interpreted as a single-server $\cdot/M/1$-type queue with service rate μ_0 and with the input which is formed by the flow of jobs rejected from the primary queue. We use notation $\cdot/M/1$ because the merged stream arriving in the orbit is not in general GI-type since it is a complex combination of a lost part of original primary customers and secondary customers returning to the orbit after unsuccessful attempts to enter the primary queue.

In this work we establish *minimal sufficient stability conditions* for the presented retrial system. Our model is quite general. To the best of our knowledge, the retrial queueing system with constant retrial rate and general renewal arrival process is considered for the first time. In [11] Fayolle has introduced a retrial system with constant retrial rate. Fayolle has derived stability conditions for the case of M/G/1/1 primary queue. In [1] Artalejo has obtained stability conditions for the Markovian case M/M/2/2. In [20] Ramalhoto and Gómez-Corral have obtained stability conditions for the M/M/1/2 case. For the general Markovian case M/M/c/K the authors of [20] have obtained decomposition results assuming ergodicity. The ergodicity conditions for the multiserver Markovian case M/M/c/c with recovery probability have been derived by Artalejo, Gómez-Corral and Neuts in [2]. In [13] Krishna Kumar and Raja have derived stability conditions for the M/M/c/c constant retrial rate model with feedback and balking. Stability conditions for the basic M/M/c/c constant retrial model can be recovered from the results in [2] and [13]. The ergodicity conditions for the Markovian case M/M/1/K have been obtained in [4]. In the works [2], [4] and [13] the authors have established stability conditions by using the matrix-analytic technique for QBD processes [19]. We shall demonstrate that stability conditions for M/G/1/1, M/M/c/c and M/M/1/K systems are particular cases of our general conditions. Furthermore, the following important cases with Poisson input have not previously been covered: M/G/c/c (Erlang model), M/M/c/K (Markovian model with a general number of servers and waiting spaces), M/D/1/K system. Stability conditions of these important cases appear to be particular cases of our general conditions and will be considered in detail below. We emphasize that, to the best of our knowledge, such retrial system with general renewal input of primary customers is considered for the first time.

There is a number of applications of retrial systems with constant retrial rates in telecommunication. Using a retrial queue with constant retrial rate Fayolle [11] has modelled a

telephone exchange system. In the series of papers [8]-[10] the authors have proposed to use a retrial queueing system with constant retrial rate to model Multiple Access protocols. In particular, in [8] the authors have modelled an unslotted Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol and in [9] and [10] the authors have modelled some versions of the ALOHA protocol. In [4] and [5] the authors have suggested to use retrial queues and retrial networks with constant retrial rates to model TCP traffic originated from short HTTP connections.

The stability analysis used in this paper is based on renewal theory and a characterization of the limiting behavior of the forward renewal time in the process generated by regenerations of a basic process. This approach, presented in a general form in [16, 18], turns out to be effective in the stability analysis of many queueing systems including general multiserver retrial queue [17], and also multiserver system with non-identical servers [15]. The presented method also works successfully outside of Markovian models, and it is demonstrated also in this paper where non-Markov processes are considered. In particular, it allows us to reduce the dimension of the processes that simplifies analysis considerably and does not require involved stability techniques developed in the theory of multi-dimensional Markov processes [6, 14]. An important contribution of this work is an extension of the stability analysis to arbitrary initial state of the system.

The paper is organized as follows. In the next Section 2 we present the main result of the paper, minimal stability conditions for GI/G/c/K-type retrial queue with constant retrial rate. We also provide the proof of the main result. Then, in Section 3 we specify the general condition for a number of important particular cases of the primary queue. In particular, we consider the general Markovian queue, the Erlang queue and the queue with deterministic service. We would like to note that stability conditions for these important particular retrial queues have not been available before. We conclude our paper with Section 4.

2 Stability analysis

To describe the behavior of the system, we consider (right-continuous) process $M(t) := N(t) + \nu(t)$, $t \geq 0$, where $N(t)$ is the number of retrial customers being in orbit and $\nu(t)$ is the number of the customers *waiting in the buffer*, at instant t. Note that $\nu(t) \in [0,\, K-c]$ for any t. Also we introduce the (right-continuous) process $W(t)$, $t \geq 0$, expressing, at each instant t, the *remaining workload in all servers*. More exactly, if $S_i(t)$ is the remaining service time at server i at instant t ($= 0$ if the server is empty), then $W(t) = \sum_{i=1}^{c} S_i(t)$. Introduce the basic (two-dimensional) process $X = \{X(t) := (M(t), W(t)), t \geq 0\}$. The choice of the basic process is motivated by the stability analysis under arbitrary initial state of the system, while there are other candidates for basic process, to analyze the system under zero initial state. In the latter case, one can use, for instance, the total number of customers in the system, or the total remaining workload in the system (including orbit), etc. Also, let $X(t_k^-) = X_k$, $k \geq 1$. Denote $T_0 = 0$, then the instants

$$T_{n+1} = \inf_k (t_k > T_n : X_k = (0, 0)), \quad n \geq 0, \tag{1}$$

are the regeneration points of the basic process X. Let T be a generic regeneration period and $T(t) = \inf_k\{T_k - t : T_k - t > 0\}$ be the forward renewal/regeneration time at instant $t \geq 0$. If

$$T_1 < \infty \text{ with probability 1 (w.p.1) and } \quad \mathsf{E}T < \infty, \tag{2}$$

then we call any regenerative process having regeneration instants defined by (1) (and also the original system) *positive recurrent*. (This term, repeatedly used in previous works [15, 16, 17, 21, 22], has evident analogy with positive recurrent Harris Markov chains whose embedded renewal process of regenerations has finite mean cycle length.)

It follows from the theory of regenerative processes that positive recurrence is the most essential element in stability analysis of the process. Indeed, if the interarrival time τ is non-lattice then T is so, and, under positive recurrence, $X(t)$ converges to a stationary limit. (It is obvious that the system is unstable if $\mathsf{E}T = \infty$.) Our approach to stability is based on the following result [12]: if $\mathsf{E}T = \infty$ then

$$T(t) \Rightarrow \infty \quad \text{as } t \to \infty, \tag{3}$$

regardless of initial (finite) delay $T_1 = T(0)$, where $\Rightarrow$ stands for convergence in probability. Thus, if convergence (3) does not hold then (2) is satisfied and positive recurrence for the *zero- delayed* process holds, in which case $T_1 = T$ and $X(0) = X_1 = (0,0)$ (that is $t_1 = 0$ is a regeneration instant). However, violation of (3) does not implies in general the finiteness of T_1 w.p.1 under arbitrary initial state $X(0)$, and an extra (sometimes hard) work is required to prove that $T_1 < \infty$. Denote $\beta_0 = 0$, then

$$\beta_{n+1} \quad = \quad \inf_k (k > \beta_n : X_k = (0,0)),\ n \geq 0, \tag{4}$$

are regeneration instants of the embedded discrete-time process X_n, $n \geq 1$, with generic regeneration cycle length β, and its remaining renewal time $\beta(n) := \inf_k\{\beta_k - n : \beta_k - n > 0\} \Rightarrow \infty$ (as $n \to \infty$) provided $\mathsf{E}\beta = \infty$.

Now we outline how to apply the above mentioned approach to our model. Because the buffer of the primary queue is finite, the source of instability of the system can only be the unlimited increase of the orbit size. Thus, we first show that under the predetermined stability condition (see Theorem 2.1) the orbit size $N(t) \not\Rightarrow \infty$. Then the second step is to show that also $T(t) \not\Rightarrow \infty$. Finally, we apply characterization (3). To establish $T_1 < \infty$ for non-zero initial state, we will use new development of the approach presented in [18].

Let us denote original system by Σ and construct an auxiliary (new) system $\hat{\Sigma}$ as follows. The system $\hat{\Sigma}$ has the same set of servers and the same buffer as system Σ, the same renewal input of primary customers with rate λ (we call them λ-customers) and in addition *an independent Poisson input of primary customers* with rate μ_0 (we call them μ_0-customers). Arriving primary customer (of any type) who finds servers and the buffer full in system $\hat{\Sigma}$ joins the orbit. The orbit is an infinite buffer system of $\cdot/M/1$ type as in the original system. The secondary customers leaving orbit (in $\hat{\Sigma}$) *leave the system forever and do not affect its future state*. Note that the system $\hat{\Sigma}$ (as Σ) regenerates at the instants when the

λ-customers find both buffer and orbit empty. Note that here we also use the memoryless property of the input of μ_0-customers. Moreover, for any variable ζ in system Σ we denote corresponding variable in system $\hat{\Sigma}$ as $\hat{\zeta}$. In particular, $\hat{W}_k$ is the remaining workload in all servers and $\hat{\nu}_k$ is the number of waiting customer in the buffer, respectively, in the system $\hat{\Sigma}$ at instant t_k^-, $k \geq 1$.

Denote the primary queue of system $\hat{\Sigma}$ as $\hat{\Theta}$, and note that it can be considered as an *isolated system* because secondary customers (in $\hat{\Sigma}$) leave the system and *do not go to the primary queue.* Then the subsystem $\hat{\Theta}$ regenerates (in continuous time) at the instants

$$\hat{\Psi}_{n+1} \quad = \quad \min(t_k > \hat{\Psi}_n : \hat{W}_k = \hat{\nu}_k = 0), \quad n \geq 0, \tag{5}$$

where, by definition, $\hat{\Psi}_0 = 0$. (So the regenerations of the whole system $\hat{\Sigma}$ are a subsequence of the regenerations of $\hat{\Theta}$.) Because the subsystem $\hat{\Theta}$ has a regenerative input and finite buffer, then such a system is positive recurrent under the condition

$$\mathsf{P}(\tau > S^{(1)}) > 0, \tag{6}$$

see [16, 18]. In particular, mean generic regeneration period $\mathsf{E}\hat{\Psi} < \infty$.

Remark 1 *Exact condition in [16] (condition (3.15) there) and in [18] (condition (31) there) require that the discrete-time regeneration period $\hat{A}$ (counting all arrivals during regeneration period $\hat{\Psi}$) equals 1 with a positive probability. It occurs if the interarrival time τ following a λ-customer (with service time $S^{(1)}$) starting new regeneration period is larger than $S^{(1)}$ and less than the next Poisson arrival (with interarrival time $\tilde{\tau}$). In other words, the following inclusion holds: $\{\hat{A} = 1\} \supseteq \{\tilde{\tau} > \hat{\tau} > S^{(1)}\}$. Obviously, under condition (6), $\mathsf{P}(\tilde{\tau} > \tau > S^{(1)}) > 0$, and the required assumption is fulfilled.*

Denote by $\hat{R}(t)$ the total number of rejected customers in system $\hat{\Theta}$ in the interval $[0, t]$ (this is also the total number of customers which went to the orbit in the whole system $\hat{\Sigma}$). Denote by $\hat{A}(t)$ the total number of arrivals (primary λ-customers and μ_0-customers) in the interval $[0, t]$. Denote also by $\hat{R}$ the number of rejected customers during regeneration cycle of system $\hat{\Theta}$.

Of course, the process $\{\hat{R}(t),\, t \geq 0\}$ is positive recurrent cumulative process with embedded regenerations $\{\hat{\Psi}_n\}$ and, in particular, there exists (w.p.1) the limit

$$\lim_{t\to\infty} \frac{\hat{R}(t)}{\hat{A}(t)} = \frac{\mathsf{E}\hat{R}}{\mathsf{E}\hat{A}}. \tag{7}$$

(To explain, we note that $\hat{R}(t)/t \to \mathsf{E}\hat{R}/\mathsf{E}\hat{\Psi}$, $\hat{A}(t)/t \to \mathsf{E}\hat{A}/\mathsf{E}\hat{\Psi}$.) In the system $\hat{\Sigma}$, define indicator I_n as

$$I_n = \begin{cases} 1, & \text{if customer } n \text{ is rejected,} \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

so the sequence $\{I_n,\, n \geq 1\}$ has regeneration period $\hat{A}$. Because $\mathsf{P}(\hat{A} = 1) > 0$, then the weak limit $I_n \Rightarrow I$ exists, or $\mathsf{P}(I_n = 1) \to \mathsf{E}I := \mathsf{P}_{loss}$ where P_{loss} is stationary loss probability.

Moreover, by the standard result of regenerative theory, stationary loss probability coincides with the long-run-average loss probability (7):

$$\mathsf{P}_{loss} = \lim_{n\to\infty} \frac{\sum_k^n I_k}{n} = \frac{\mathsf{E}\hat{R}}{\mathsf{E}\hat{A}}. \tag{9}$$

Now we are ready to formulate the main stability result.

Theorem 1 *Assume that condition (6) and the following condition*

$$(\lambda + \mu_0)\mathsf{P}_{loss} < \mu_0, \tag{10}$$

hold. Then, under arbitrary (fixed) initial state $X(0) = X_1 = (m_0,\, w_0)$, the original system is positive recurrent, that is

$$\mathsf{E}T < \infty,\, \mathsf{E}\beta < \infty \;\; \textit{and} \;\; T_1 < \infty, \;\; \beta_1 < \infty \;\; \textit{w. p. 1.}$$

Moreover, the stationary distribution $\lim_{n\to\infty} \mathsf{P}(X_n \in \cdot)$ exists. If, in addition, interarrival time τ is non-lattice, then stationary distribution $\lim_{t\to\infty} \mathsf{P}(X(t) \in \cdot)$ also exists.

Proof. We will use a monotonicity property of the loss system with respect to change of service times, see [24, 25, 26]. Namely, we use coupling to sample identical corresponding interarrival times in both systems Σ, $\hat{\Sigma}$ until first empty period of the orbit in the original system Σ. Then we assume that during this (and any following) empty period of the orbit we continue to sample input Poisson process (with parameter μ_0) in the original system but with arriving customers having *zero service times* (as long as the empty period lasts). At the end of such period, we interrupt current interarrival times (in Poisson processes in both systems Σ and $\hat{\Sigma}$) and resample new (identical) interarrival times for both systems. By memoryless property, this resampling keeps distribution of the input process. At the same time this procedure allows us to keep equivalence between input intervals in both systems. (Note that λ-inputs stay identical and unchanged in both systems.) Moreover, this shows that service times of arriving customers in the systems are stochastically ordered as $S_n^{(1)} \leq_{st} \hat{S}_n^{(1)}$ while for non-zero (actual) customer n, $S_n^{(1)} =_{st} \hat{S}_n^{(1)}$, $n \geq 1$.

Now we can use the result of [24] which claims (in adaptation to our model) that if two (the same) finite capacity systems have the same input and the ordered service times, as above, then the number of rejected customers in primary queues of systems Σ and $\hat{\Sigma}$, respectively, in the interval $[0,\, t]$ are ordered as

$$R(t) \leq_{st} \hat{R}(t), \;\; t \geq 0. \tag{11}$$

We emphasize that $R(t)$ counts all rejections which happen in system Σ including repetitive rejections of the orbit customers after unsuccessful attempts to enter the primary queue. Because the input in system $\hat{\Sigma}$ is a superposition of two independent renewal processes, then w.p. 1 as $t \to \infty$

$$\frac{A(t)}{t} \to \lambda + \mu_0. \tag{12}$$

Since $\hat{R}(t)=\sum_k^{A(t)} I_k$, it then follows from (7) and (9) that w.p.1

$$\lim_{t\to\infty}\frac{\hat{R}(t)}{t}=(\lambda+\mu_0)\mathsf{P}_{loss}. \tag{13}$$

(It follows from $\hat{R}(t)\le A(t)$ and $\mathsf{E}A(t)/t\to\lambda+\mu_0$, that the family $\{\hat{R}(t)/t,\, t\ge 0\}$ is uniformly integrable, and thus the limit $\mathsf{E}\hat{R}(t)/t\to(\lambda+\mu_0)\mathsf{P}_{loss}$ also exists.)

Denote by $V_o(t)$, $\hat{V}_o(t)$ the total workloads arrived *to the orbit* in the systems Σ, $\hat{\Sigma}$, respectively, during the interval $[0,\, t]$, including the same (arbitrary) initial state $V_o(0)=\hat{V}_o(0):=V_o$. Note that

$$\hat{V}_o(t)=V_o(0)+\sum_{k=1}^{\hat{R}(t)}S_k^{(2)},\;\; t\ge 0 \quad (\sum_{\emptyset}=0), \tag{14}$$

and thus,

$$\frac{\hat{V}_o(t)}{\hat{R}(t)}\to\frac{1}{\mu_0}\;\text{ w.p.1 as } t\to\infty. \tag{15}$$

Denote, in the original system Σ, by $\mu_o(t)$ the total empty time of the orbit in interval $[0,\, t]$ and by $W_o(t)$ the (right-continuous) remaining workload in orbit at instant $t\ge 0$. Now we have the following balance equation

$$V_o(t)=W_o(t)+t-\mu_o(t),\;\; t\ge 0. \tag{16}$$

Thus, we have

$$\begin{aligned}\mu_o(t) \quad &\ge \quad t-V_o(t)=t-\sum_{k=1}^{R(t)}S_k^{(2)}-V_o(0)\\ &\ge_{st} \quad t-\sum_{k=1}^{\hat{R}(t)}S_k^{(2)}-\hat{V}_o(0)=t-\hat{V}_o(t),\;\; t\ge 0.\end{aligned} \tag{17}$$

By (10), (13), (15) this implies

$$\liminf_{t\to\infty}\frac{\mu_o(t)}{t}\ge 1-\frac{(\lambda+\mu_0)\mathsf{P}_{loss}}{\mu_0}:=\delta_0>0. \tag{18}$$

This also shows that under arbitrary initial state $V_o(0)$,

$$\mu_o(t)=\int_0^t I(N(u)=0)du\to\infty\;\text{ w.p.1 as }\; t\to\infty, \tag{19}$$

where I denotes indicator function. By Fatou's lemma,

$$\liminf_{t\to\infty} \frac{1}{t}\mathsf{E}\mu_o(t) > 0. \tag{20}$$

Thus, $\mathsf{P}(N(t) = 0) \not\to 0$ as $t \to \infty$ and there exist $\delta^* > 0$ and non-random sequence of instants $z_n \to \infty$ such that

$$\inf_{n\geq 1} \mathsf{P}(N(z_n) = 0) \geq \delta^*. \tag{21}$$

Starting from this point we work with the original system Σ only. Let $\tau(t) = \min_k(t_k - t : t_k - t \geq 0)$ be the remaining interarrival time at instant t in system Σ. Note that, for each z_n,

$$\mathsf{P}(N(z_n) = 0) = \mathsf{P}\Big(N(z_n + \tau(z_n)) = 0\Big) \geq \delta^*. \tag{22}$$

Denote $\kappa(n) = \min(k : t_k \geq z_n)$, then $z_n + \tau(z_n) = t_{\kappa(n)}$ is the first arrival instant after z_n, $n \geq 1$. Let $L(t)$ be the total remaining workload at the servers and buffer at instant t. Obviously,

$$L(t) \leq_{st} W(t) + \sum_{i=1}^{K-c} S_i^{(1)},\ t \geq 0. \tag{23}$$

Note that the remaining service time processes, $\{S_i(t),\, t \geq 0\}$, $i = 1,\dots,c$, are tight. (To have the tightness, the only requirement, besides the finiteness of $\mathsf{E}S^{(1)}$, is that service times $\{S_n^{(1)}\}$ are independent of the input process [15].) Thus, the process $W(t)$, $t \geq 0$, is tight, and hence, the process $L(t)$, $t \geq 0$, is also tight. Then we can find a constant D such that $\inf_n \mathsf{P}(L(z_n) \leq D) \geq 1 - \delta^*/2$. Now we fix for a moment some z_n satisfying (21) and let $L(t_n^-) = L_n$, $n \geq 1$. Then, we have

$$\mathsf{P}(N_{\kappa(n)} = 0,\, L_{\kappa(n)} \leq D) \geq \mathsf{P}(N(z_n) = 0,\, L(z_n) \leq D) \geq \frac{\delta^*}{2}, \tag{24}$$

where the first inequality is valid since the workload decreases in interval $[z_n,\, t_{\kappa(n)})$ and the second inequality follows from

$$\begin{aligned}\mathsf{P}(N(z_n) = 0,\, L(z_n) \leq D) &= \mathsf{P}(L(z_n) \leq D) - \mathsf{P}(N(z_n) > 0,\, L(z_n) \leq D)\\ &\geq \mathsf{P}(L(z_n) \leq D) - \mathsf{P}(N(z_n) > 0).\end{aligned}$$

Next we introduce the event

$$\mathcal{E}_n := \{N_{\kappa(n)} = 0,\, L_{\kappa(n)} \leq D,\, \tau(z_n) \leq C\}$$

where, by the tightness of the process $\{\tau(t), t \geq 0\}$, the constant C is taken in such a way that $\mathsf{P}(\mathcal{E}_n) \geq \delta^*/4$. Thus, on the event $\mathcal{E}_n$, the customer $\kappa(n)$ indeed arrives in interval

$[z_n,\, z_n + C]$ and finds the workload $L_{\kappa(n)} \le D$. It follows from (6) and $\mathsf{E}\tau < \infty$ that one can find constants $a < \infty$, $\gamma > 0$, $\epsilon > 0$ such that

$$\mathsf{P}(a \ge \tau > \gamma + S^{(1)}) = \epsilon. \tag{25}$$

Note that, for each i, on the event $\omega_i := \{a \ge \tau_i > \gamma + S_i^{(1)}\}$ the workload accumulated in the system at instant t_i, decreases during interarrival time $[t_i,\, t_{i+1})$ not less than by γ, *provided that at least one server is not empty* during this interval. If only orbit is not empty, then an orbit customer may attempt to enter server/buffer before next arrival instant t_{i+1}. In such a case, regardless of the retrial attempt was successful or not, on the event ω_i, the accumulated workload decreases not less than by $\gamma/2$ (during $[t_i,\, t_{i+1})$) with the probability $\ge \epsilon(1 - e^{-\mu_0 \frac{\gamma}{2}})$. Denote $R := \lceil 2D/\gamma \rceil$. Then on the event

$$\mathcal{E}_n \bigcap \bigcap_{i=\kappa(n)}^{\kappa(n)+R} \omega_i$$

R primary customers arrive in the interval $[z_n,\, z_n + C + aR]$, and the accumulated workload reaches zero (regeneration) with probability $\ge \frac{1}{4}\delta^* \epsilon^R (1 - e^{-\mu_0 \frac{\gamma}{2}})^R > 0$. Because this lower bound is uniform in n, z_n, then $\mathsf{E}\beta < \infty$. It follows from Wald's identity and representation $T =_{st} \sum_{k=1}^{\beta} \tau_k$ that mean cycle period for continuous time processes is finite, $\mathsf{E}T = \mathsf{E}\beta \mathsf{E}\tau < \infty$.

Using a modification of the approach from [18], we now extend analysis to arbitrary initial state of the basic process X. Let $S_{k(i)}^{(1)}$ be the k-th service time realized at server i. (For each i, these service times are i.i.d. and distributed as $S^{(1)}$.) Introduce

$$\tilde{S}_i(t) = \min_{k(i)} (S_{1(i)}^{(1)} + \cdots + S_{k(i)}^{(1)} - t : \; S_{1(i)}^{(1)} + \cdots + S_{k(i)}^{(1)} - t > 0)$$

the remaining renewal time at instant t in the renewal process generated by service times of server i, and let $\tilde{S}_i(0) = S_i(0)$, $i = 1, \ldots, c$. For any (integer) $r \ge 0$ and $x \ge 0$, denote the set $\mathsf{B}(r,\, x) = [0,\, r] \times [0,\, x]$ and consider the process $Y =: \{Y(t) = (N(t),\, W(t)),\, t \ge 0\}$. Then, for any t and x:

$$\begin{aligned} I(Y(t) \in \mathsf{B}(0,\, x)) &\ge I(N(t) = 0) - I(W(t) > x) \\ &\ge I(N(t) = 0) - \sum_{i=1}^{c} I\Big(S_i(t) > \frac{x}{c}\Big). \end{aligned} \tag{26}$$

Although, for each t, $S_i(t)$ and $\tilde{S}_i(t)$ in general are not comparable, but by construction, the following inequality holds for any x:

$$\int_0^t I\Big(S_i(u) > x\Big) du \le \int_0^t I\Big(\tilde{S}_i(u) > x\Big) du, \; i = 1, \ldots, c. \tag{27}$$

Note that $I\Big(\tilde{S}_i(t) > x\Big)$, $t \geq 0$, is a regenerative process with (generic) cycle length $S^{(1)}$. Then w.p.1 as $t \to \infty$,

$$\frac{1}{t}\int_0^t I\Big(\tilde{S}_i(t) > x\Big) \to \frac{1}{\mathsf{E}S^{(1)}}\int_x^\infty \mathsf{P}(S^{(1)} > t)dt := \mathsf{P}(S_e^{(1)} > x), \;\; i = 1, \dots, c, \tag{28}$$

where distribution of the stationary overshoot $S_e^{(1)}$ is proper. Hence, we can take x_1 in such a way that

$$\mathsf{P}\Big(S_e^{(1)} > \frac{x_1}{c}\Big) < \frac{\delta_0}{2c}. \tag{29}$$

If we take $x = x_1$ in (26)-(35) then it follows from (18), (26)–(29) that w.p.1

$$\liminf_{t\to\infty} \frac{1}{t}\int_0^t I(Y(u) \in \mathsf{B}(0,\, x_1))du \geq \liminf_{t\to\infty} \frac{\mu_o(t)}{t} - \frac{\delta_0}{2} = \frac{\delta_0}{2}. \tag{30}$$

In particular, the total time that the process Y spends in the set $\mathsf{B}(0,\, x_1)$ within time interval $[0,\, t]$

$$\int_0^t I(Y(u) \in \mathsf{B}(0,\, x_1))du \to \infty, \quad t \to \infty. \tag{31}$$

Because, for each t and $x \geq 0$,

$$\{Y(t) \in \mathsf{B}(0,\, x)\} \subseteq \{X(t) \in \mathsf{B}(K-c,\, x)\}, \tag{32}$$

then, by (31), the total time the process X spends in the set $\mathsf{B}(K-c,\, x_1)$ during time interval $[0,\, t]$

$$\mu_B(t) := \int_0^t I(X(u) \in \mathsf{B}(K-c,\, x_1))du \to \infty \;\text{ w.p.1 as }\; t \to \infty.$$

Denote by $\Lambda(t)$ the number of λ-customers arriving in the interval $[0,\, t]$, and let $G_B(t)$ be the number of these λ-customers who meets the process X in the set $\mathsf{B}(K-c,\, x_1)$, that is

$$G_B(t) = \sum_{k=1}^{\Lambda(t)} I(X_k \in \mathsf{B}(K-c,\, x_1)). \tag{33}$$

It then follows that w.p.1

$$G_B(t)(\max_{1\leq n\leq \Lambda(t)} \tau_n) \geq \mu_B(t) \to \infty,\; t \to \infty. \tag{34}$$

Since $\mathsf{E}\tau < \infty$, then $\max_{1\leq n\leq \Lambda(t)} \tau_n = o(t),\;\; t \to \infty$ [23]. Thus, $G_B(t) \to \infty$ w.p. 1 as $t \to \infty$, and the total number of λ-customers, $G_B(\infty)$, meeting the process X in the set

$\mathtt{B}(K-c,\, x_1)$ is infinite w.p. 1. (This holds for arbitrary initial state $X(0)$, see (18),(19),(30).) Recall initial state $X(0)=(m_0,\, w_0)$, and choose and fix arbitrary $r \geq \max(m_0,\, K-c)$ and $z \geq \max(w_0,\, x_1)$. Then, as in [18], one can find constants $D_0 < \infty$, $\varepsilon_0 > 0$ such that the remaining renewal time $\beta(k)$ satisfies

$$\inf_{k\geq 1} \mathsf{P}_{(m_0,\, w_0)}(\beta(k) \leq D_0 \,|\, X_k \in \mathtt{B}(r,\, z),\, \beta_1 > k) \geq \varepsilon_0, \tag{35}$$

where the bounds ε_0 and D_0 are uniform: they depend on r, z but does not depend on neither customer number k nor the specific states $X_k \in \mathtt{B}(r,\, z)$. Indeed, due to the specific form of the basic process X, conditioned on the event $\{X_k \in \mathtt{B}(r,\, z),\, \beta_1 > k\}$, the total workload in the system at instant t_k^- is (uniformly in k) stochastically upper bounded by the quantity

$$z + \sum_{i=1}^{K-c} S_i^{(1)} + \sum_{i=1}^{r} S_i^{(2)}, \tag{36}$$

where, evidently, both the service times $\{S_i^{(1)}\}$ of waiting customers in the buffer and (exponential) retrial times $\{S_i^{(2)}\}$ of all orbit customers are i.i.d and independent of the remaining workload $W_k\, (\leq z)$ in the servers. Then one can unload the system during D_0 λ-arrivals with a probability $\geq \varepsilon_0$ (like in the proof after formula (30)), and thus condition (35) is indeed satisfied. Then one can show that the mean number of λ-customers meeting the process X in the set $\mathtt{B}(r,\, z)$ during first regeneration period is finite,

$$\mathsf{E}_{(m_0,\, w_0)}\Big(\sum_{k=1}^{\beta} I(X_k \in \mathtt{B}(r,\, z))\Big) \leq \frac{D_0}{\varepsilon_0} < \infty. \tag{37}$$

(More details can be found in [17, 18].) Because $\mathtt{B}(r,\, z) \supseteq \mathtt{B}(K-c,\, x_1)$, then (37) implies that the number of λ-customers meeting the process X in the set $\mathtt{B}(K-c,\, x_1)$ during first regeneration period is finite, $G_B^{(1)} := \sum_{k=1}^{\beta} I(X_k \in \mathtt{B}(K-c,\, x_1)) < \infty$. Then it follows from above that $\mathsf{P}_{(m_0,\, w_0)}(G_B^{(1)} < G_B(\infty)) = 1$, and hence the number of regeneration cycles is not less than two. Thus, first regeneration period $\beta_1 < \infty$ and hence $T_1 = \tau_1 + \cdots + \tau_{\beta_1} < \infty$. (Indeed, the number of regeneration cycles is infinite w.p.1.) Finally, we note that condition (6) implies aperiodicity of β. ■

Remark 2 *In general, one can define another type of regeneration, for instance, k-regeneration, when an arriving λ-type customer see all servers and buffer empty and a fixed k customers in the orbit, see [17]. Moreover, a relaxation (or elimination) of assumption (6) (which is necessary to have classical regenerations of type (1), (4)) may lead to the so-called one-dependent (or weak) regeneration with a dependence between adjacent regeneration cycles, see for instance, [16, 21, 22]. However, we do not consider such regenerations in this work.*

Remark 3 *In a quite general case of unbounded τ (and in particular for the Poisson λ-arrival process), the constant a in (26) can be taken in such a way that regeneration occurs in the interval $[z_n, z_n + C + a]$.*

3 Application to particular queueing models

In this section we apply general stability conditions of Theorem 1 to important particular types of the primary queue.

3.1 Markovian case M/M/c/K

In the Markovian case with c servers, service rate μ and K places in the primary queueing system, the loss probability in the auxiliary Σ_2 system is given by

$$P_{loss} = \frac{((\lambda+\mu_0)/\mu)^c}{c!}\left(\frac{(\lambda+\mu_0)/\mu}{c}\right)^{K-c} P_0,$$

where

$$P_0 = \left[\sum_{n=0}^{c} \frac{((\lambda+\mu_0)/\mu)^n}{n!} + \frac{((\lambda+\mu_0)/\mu)^c}{c!}\sum_{n=1}^{K-c}\left(\frac{(\lambda+\mu_0)/\mu}{c}\right)^n\right]^{-1}.$$

Hence, the stability condition (10) takes the form

$$\frac{((\lambda+\mu_0)/\mu)^K}{c!}\left(\frac{(\lambda+\mu_0)/\mu}{c}\right)^{K-c} P_0 < \frac{\mu_0}{\lambda+\mu_0}. \tag{38}$$

The other conditions of Theorem 1 are naturally satisfied by the Poissonian arrival process. In the case of one server (M/M/1/K case), the above condition reduces to

$$((\lambda+\mu_0)/\mu)^K \left[\sum_{n=0}^{K} ((\lambda+\mu_0)/\mu)^n\right]^{-1} < \frac{\mu_0}{\lambda+\mu_0},$$

which is equivalent to the ergodicity condition provided in [4]. Then, in the M/M/c/c case, the condition (38) reduces to

$$\frac{((\lambda+\mu_0)/\mu)^c}{c!}\left[\sum_{n=0}^{c} \frac{((\lambda+\mu_0)/\mu)^n}{n!}\right]^{-1} < \frac{\mu_0}{\lambda+\mu_0}, \tag{39}$$

which is equivalent to the ergodicity condition provided in [2] with the recovery probability $p = 1$.

It is important to note that the conditions given above are in fact necessary and sufficient stability conditions for the Markovian case. This is why we call conditions presented in Theorem 1 minimal sufficient stability conditions.

3.2 Erlang model

When the primary queue is described by the Erlang model (M/G/c/c case), the general stability condition (10) takes the form

$$\frac{\Big((\lambda+\mu_0)\mathsf{E}S^{(1)}\Big)^c}{c!}\left[\sum_{n=0}^{c}\frac{\Big((\lambda+\mu_0)\mathsf{E}S^{(1)}\Big)^n}{n!}\right]^{-1} < \frac{\mu_0}{\lambda+\mu_0},$$

which is the same form as in (39). Thus, the stability conditions demonstrate insensitivity property when the primary queue is the Erlang queue. We would like to emphasize that it seems to be very difficult (if possible at all) to obtain the above stability condition using stability techniques for Markov chains and embedded Markov chains. Anyway, the presented approach ensures short and simple proof of stability.

3.3 Deterministic service

We can obtain explicit stability conditions in another important case, in the case of deterministic service (M/D/1/K model for the primary queue). In the case of deterministic service, the loss probability is given by

$$P_{loss} = 1 - \frac{b_{K-1}}{1+\rho b_{K-1}},$$

where

$$b_{K-1} = \sum_{n=0}^{K-1}\frac{(-1)^n}{n!}(K-1-n)^n\rho^n e^{(K-1-n)\rho},$$

with $\rho = \lambda S$ [7]. Thus, in the M/D/1/K case the stability condition (10) takes the form

$$\frac{\lambda}{\lambda+\mu_0} < \frac{b_{K-1}}{1+\rho b_{K-1}},$$

and we note that there exists an efficient recursive approach to calculate the coefficient b_{K-1} [7].

4 Conclusion

We have considered a retrial queueing system with general renewal arrival process, general service time and constant retrial rate. To the best of our knowledge a retrial queueing system with constant retrial rate was considered for the first time under so general setting. We have obtained minimal stability conditions which are necessary and sufficient in the Markovian case. Stability analysis also covers arbitrary initial state of the system. The conditions have clear probabilistic interpretation and can be easily applied to a number of important

particular cases. Examples of such particular cases are Erlang model and deterministic service model. We have observed that the stability condition for the Erlang model with retrial is insensitive to the distribution shape.

Acknowledgements

We would like to acknowledge the financial support provided by EGIDE ECO-NET grant no.18933SL and RFBR grant no.10-07-00017.

References

[1] J.R. Artalejo, "Stationary analysis of the characteristics of the M/M/2 queue with constant repeated attempts", *Opsearch*, v.33, pp.83-95, 1996.

[2] J.R. Artalejo, A. Gómez-Corral and M.F. Neuts, "Analysis of multiserver queues with constant retrial rate", *European Journal of Operational Research*, v.135, pp.569-581, 2001.

[3] S. Asmussen, *Applied Probability and Queues* (2nd ed.). Springer-Verlag, NY, 2003.

[4] K. Avrachenkov and U. Yechiali, "Retrial networks with finite buffers and their application to Internet data traffic", *Probability in the Engineering and Informational Sciences*, v.22, pp.519-536, 2008.

[5] K. Avrachenkov and U. Yechiali, "On tandem blocking queues with a common retrial queue", *Computers and Operations Research*, v.37, no.7, pp.1174-1180, 2010.

[6] A.A. Borovkov, *Ergodicity and stability of stochastic processes*, Wiley, 1998.

[7] O. Brun and J.-M. Garcia, "Analytical solution of finite capacity M/D/1 queues", *Journal of Applied Probability*, v.37, pp.1092-1098, 2000.

[8] B.D. Choi, Y.W. Shin and W.C. Ahn, "Retrial queues with collision arising from unslotted CSMA/CD protocol", *Queueing Systems*, v.11, pp.335-356, 1992.

[9] B.D. Choi, K.K. Park and C.E.M. Pearce, "An M/M/1 retrial queue with control policy and general retrial times", *Queueing Systems*, v.14, pp.275-292, 1993.

[10] B.D. Choi, K.H. Rhee and K.K. Park, "The M/G/1 retrial queue with retrial rate control policy", *Probability in the Engineering and Informational Sciences*, v.7, pp.29-46, 1993.

[11] G. Fayolle, "A simple telephone exchange with delayed feedback", In O.J. Boxma, J.W. Cohen and H.C. Tijms (eds.), *Teletraffic Analysis and Computer Performance Evaluation*, v.7, pp.245-253, Elsevier, Amsterdam, 1986.

[12] W. Feller, *An Introduction to Probability Theory and its Applications. I* (2nd ed.), Wiley, NY, 1971.

[13] B. Krishna Kumar and J. Raja, "On multiserver feedback retrial queues with balking and control retrial rate", *Annals of Operations Research*, v. 141, pp.211-232, 2006.

[14] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[15] E. Morozov, "The tightness in the ergodic analysis of regenerative queueing processes", *Queueing Systems*, v.27, pp.179-203, 1997.

[16] E. Morozov, "Weak regeneration in modeling of queueing processes", *Queueing Systems*, v.46, pp.295-315, 2004.

[17] E. Morozov, "A multiserver retrial queue: Regenerative stability analysis", *Queueing Systems*, v.56, pp.157-168, 2007.

[18] E. Morozov and R. Delgado, "Stability analysis of regenerative queues", *Automation and Remote control*, v.70, no.12, pp.1977-1991, 2009.

[19] M.F. Neuts, *Matrix-geometric solutions in stochastic models – An algorithmic approach*, John Hopkins University Press, Baltimore, 1981.

[20] M.F. Ramalhoto and A. Gómez-Corral, "Some decomposition formulae for M/M/r/r+d queues with constant retrial rate", *Stochastic Models*, v.14, pp.123-145, 1998.

[21] K. Sigman, "One-dependent regenerative processes and queues in continuous time", *Math. Oper. Res.*, v.15, pp.175-189, 1990.

[22] K. Sigman and R.W. Wolff, "A review of regenerative processes", *SIAM Review*, v.35, pp.269-288, 1993.

[23] W.L. Smith, "Regenerative stochastic processes", *Proc. Royal Soc.*, Ser. A, v.232, pp.6-31, 1955.

[24] D. Sonderman, "Comparing multi-server queues with finite waitng rooms, I: Same number of servers", *Advances in Applied Probability*, v.11, pp.439-447, 1979.

[25] D. Sonderman, "Comparing multi-server queues with finite waitng rooms, II: Different number of servers", *Advances in Applied Probability*, v.11, pp.448-455, 1979.

[26] W. Whitt, "Comparing counting processes and queues", *Advances in Applied Probability*, v.13, pp.207-220, 1981.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399

INRIA

LORRAINE

INRIA

Rapport technique